VAWKUM Transactions on Computer Sciences
http://vfast.org/journals/index.php/VTCS@ 2016, ISSN(e):2308-8168, ISSN(p): 2411-6335
Volume 11, Number 1, September-October , 2016 pp. **01—09**

WEB USERS CLUSTERING BASED ON FUZZY C-MEANS

WALEED ALI *[1] AND MOHAMMED ALRABIGHI [1]

[1] Department of Information Technology, Faculty of Computing and Information Technology, King Abdulaziz University, Rabigh, Kingdom of Saudi Arabia
*Email: waleedalodini@gmail.com

Revised September 2016

ABSTRACT. *The Web contributes greatly to our life in many fields such as education, entertainment, Internet banking, online shopping and software downloading. This has led to rapid growth in the number of Internet users, which resulting in an explosive increase in traffic or bottleneck over the Internet performance. This paper proposes a new approach to group users according to their Web access patterns. The proposed approach for grouping users is based on Fuzzy c-means technique, which allows web users to be assigned into more than one cluster or interest. Each web user has a degree of membership of belonging to each cluster. The experimental results showed that the web users were successfully clustered to similar groups very fast using Fuzzy-c-means. In addition, the Fuzzy-c-means performed well and became much better when the clusters number increased on two real Bo2 and NY datasets. The proposed intelligent web users clustering based on Fuzzy-c-means can be used for discovering users' interests in Web pages that can contribute in enhancing several approaches such as Web caching, Web pre-fetching and Web recommender systems that are recently used to improve the Web performance.*

Keywords*:* Web users; Clustering; Fuzzy-c-means technique.

1. **Introduction.** The World Wide Web (Web) is the most common and significant service on the Internet, which play key roles in many real applications such as education, entertainment, Internet banking, online shopping and software downloading. The Web has become the most popular medium, which enables massive information publishing and retrieval. However, the explosive growth of the Web has led to an information overload that continuously expands causing various problems to web users. These problems are usually related to the accuracy of the retrieved information, which is characterized by low precision or irrelevance. Moreover, the delivered Web information lacks personalization while the requested web pages do not conform with users preferences [1, 2, 3].

Web user clustering is one of the most effective web mining tasks, which has given solution to the above limitations. The web user clustering can create clusters or groups of users with similar browsing interests and pages. Accordingly, Web designers can analyze the characteristics of the clusters in order to understand the common user's preferences better and may provide more suitable, customized services to the users [4, 5, 6, 7].

Several research works have suggested intelligent clustering approach for discovering users' interests in Web pages. ART Neural Networks were proposed for Web user clustering and then utilized in web prefetching [9, 10]. K-means algorithm has been used by [6, 7] in order to cluster the Web users to similar groups based on users' interests in Web pages. Chimphlee *et al.* [10] presented a rough set clustering to cluster web transactions from web access logs and used Markov model for next access prediction.

Although the web users may like many web pages with different interests, the existing Web user clustering approaches were assigned each web user to just one cluster or interest. In this paper, the Web users

clustering is carried out based on Fuzzy c-means, which allows web users to be assigned into more than one cluster or interest. Each web user has a degree of membership (or probability) of belonging to each cluster. The analysis of the Web users clustering based on Fuzzy c-means can be useful for inferring user statistics in order to improve various Web applications such as Web caching and prefetching, Web recommender system , market segmentation in e-commerce applications, and personalized Web content for users.

The remaining parts of this paper are organized as follows. Background and related works are presented in Sections 2 and 3. Web user clustering is presented in Sections 2, while Section 3 describes briefly Fuzzy-c-means clustering used in this study. In Section 4, a methodology of Web users clustering based on Fuzzy C-Means is explained. Section 5 presents and discusses results of Web users clustering based on Fuzzy C-Means. Finally, Section 6 summarizes and concludes the works presented in this paper.

2. Web User Clustering. Web mining is the use of data mining techniques to automatically discover and extract information from the Web documents and services. There are three areas of the Web mining: Web content mining, Web structure mining and Web usage mining. The Web content mining focuses on the discovery of useful information from the Web contents. The Web structure mining attempts to discover the model underlying the link structure of the Web. The Web usage mining attempts to discover knowledge for the data generated by the Web surfer's sessions or behaviors [11].

In the Web usage mining, Web documents and users clustering is one of the most interesting tasks, which can contribute to many applications such as Business Intelligence, E-Commerce, Customer Relationship Management, User Profiling and Personalization, Web site Construction, Fraud Detection, and Web Cache Replacement and Prediction [4, 5, 6, 7]. As shown in **Figure-1**, the Web user clustering is grouping of clusters of users exhibiting similar browsing patterns, while The Web document clustering is the establishment of documents with related content.

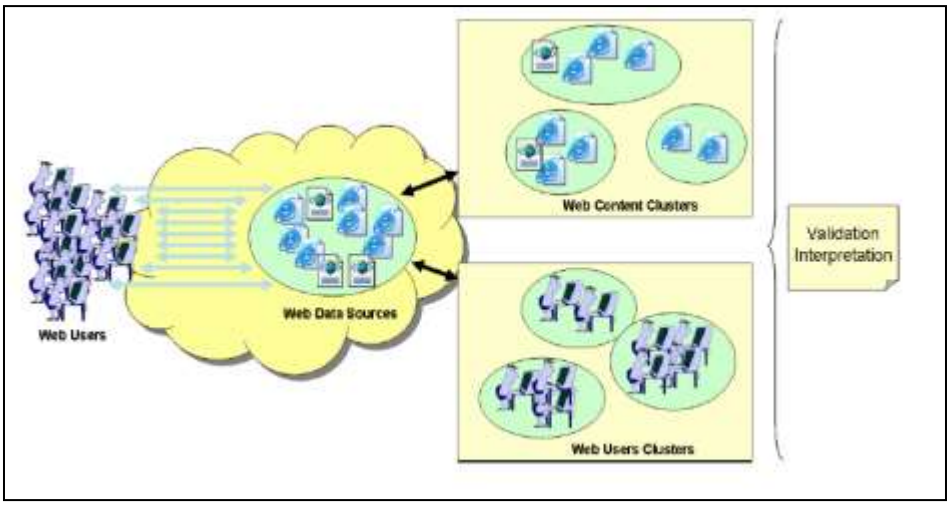

Figure-1:.Web documents and users clustering

The Web users usually perform their interest-oriented actions by clicking or visiting Web pages, which are traced in access log files. Web access logs serve as a substantial source of information about users' Web access patterns. Therefore, the Web access logs can be used to analyze and discover useful information about users' interests with the site [9, 10].

Clustering Web user access patterns may capture common user interests to a Web site, and in turn, build user profiles for advanced Web applications, such as Web caching and prefetching. The main issue of web users clustering is to use web access log files to partition a set of users into clusters such that the users within a cluster are more similar to each other than users from different clusters.

3. Fuzzy-C-Means Clustering. Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis,

information retrieval, bioinformatics, data compression, and computer graphics.

K-means clustering is one of the popular unsupervised learning algorithms that are used successfully in a wide range of applications. In K-means or Hard C-means, K cluster centers are defined as the means or centroids of the points in the cluster. Every point is assigned to one of the K clusters. The objective in K-means is to minimize the average distance of points from their cluster centers.

Fuzzy C-Means clustering method was introduced by Bezdek [12], as extension of Hard C-Mean clustering method. The Fuzzy C-Means is an unsupervised clustering algorithm that is applied to wide range of problems connected with feature analysis, clustering and classifier design [13, 14].

Unlike Hard C-Mean clustering method, fuzzy clustering provides a degree of membership of each point for every cluster. Fuzzy C-means (FCM) is the most popular fuzzy clustering algorithm. The Fuzzy C-means works as K-means except that it produces a membership matrix, which contains the degree of membership of a point to all the clusters. The Fuzzy C-means (FCM) aims to minimize the following objective function:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2 \quad , \quad 1 \le \mathrm{m} < \infty \qquad (1)$$

where m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the ith of d-dimensional measured data, c_j is the is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center .

The Fuzzy C-means Fuzzy clustering is achieved by iteratively updating of membership u_{ij} and the cluster centers c_j as shown in equations (2) and (3) in order to optimize the objective function shown above. The algorithm of the Fuzzy C-means is shown in **Figure-2**.

1. Initialize $U=[u_{ij}]$ matrix, $U^{(0)}$
2. At k-step: calculate the centers vectors $C^{(k)}=[c_j]$ with $U^{(k)}$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

3. Update $U^{(k)}$, $U^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|} \right)^{\frac{2}{(m-1)}}}$$

4. If $\| U^{(k+1)} - U^{(k)} \| < \varepsilon$ then STOP; otherwise return to step 2.

Figure-2: The algorithm of the Fuzzy C-means

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{\ m} x_i}{\sum_{i=1}^{N} u_{ij}^{\ m}} \qquad (2)$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|} \right)^{\frac{2}{(m-1)}}} \qquad (3)$$

4. A Methodology of Web Users Clustering Based on Fuzzy C-Means. The methodology of the proposed Web users clustering method based on Fuzzy c-means is illustrated in **Figure-3**. As shown in **Figure-3**, the methodology is achieved through four phases: dataset collection from proxy server, feature extraction, preparation of training dataset, and Web users clustering using fuzzy-c-means.

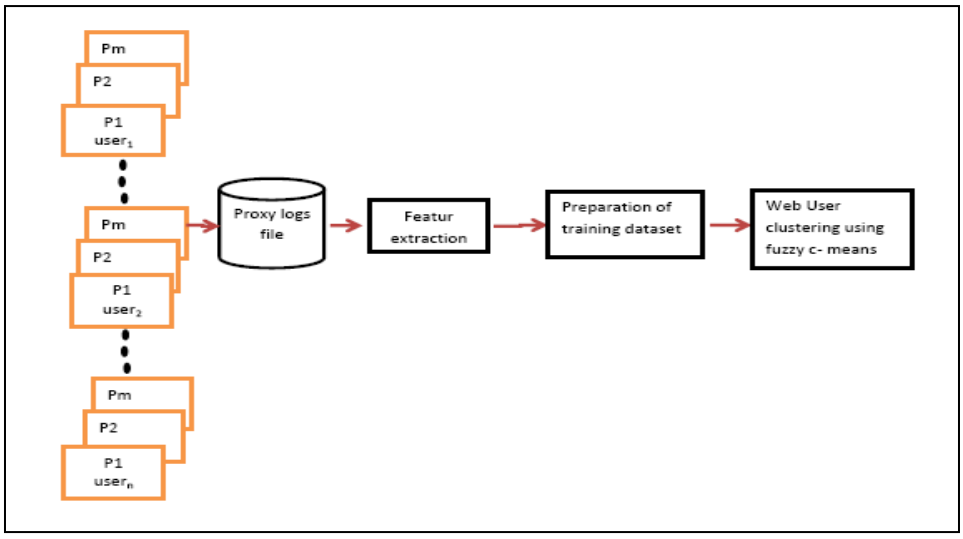

Figure-3: A methodology of Web users clustering method based on Fuzzy c-means

4.1. Data Collection. A proxy server is an intermediate server, which situated between clients and web server. When the clients request many web pages from web servers via a web proxy server, the proxy server maintains the entries to the logs as a separate log file for gathering the information of the users. In the proxy servers, information about the behaviors of groups of users in accessing a huge number of web servers are recorded in files known as web proxy logs files as shown **Figure-4**. The web proxy logs files provide information about activities performed by the users when the users log onto the servers. The proxy logs files can be obtained from proxy servers located in various organizations or universities. The web proxy logs files are considered a complete and prior knowledge and can be utilized as training data for clustering ans discovering the users' interests. **Figure-4** shows a sample of proxy logs file, which is used for clustering the users based on visited pages.

1282348821.049 138 132.55.200.134 TCP_MISS/301 471 GET http://dishnetwork.com/ - DIRECT/205.172.147.51 text/html
1282348917.977 286 249.78.126.183 TCP_MISS/302 1753 GET http://www.hotmail.com/ - DIRECT/64.4.20.184 text/html
1282348954.666 393 132.55.200.134 TCP_MISS/200 22321 GET http://www.msn.com/ - DIRECT/65.55.17.26 text/html
1282348982.398 200 249.78.126.183 TCP_MISS/200 6550 GET http://www.btt.com.ar/foto/t/12/75/1275530489_mark-webb2.jpg - DIRECT/72.232.178.138 image/jpeg
1282348982.437 286 249.78.126.183 TCP_MISS/200 22520 GET http://www.btt.com.ar/foto/t/12/81/1281356302_DSC03379.JPG - DIRECT/72.232.178.138 image/jpeg
1282349600.500 78 50.83.47.141 TCP_MISS/200 344 POST http://app.ninjasaga.com/amf/ - DIRECT/75.126.166.176 application/x-amf
1282361594.519 604 171.11.238.157 TCP_MISS/200 2209 GET http://nt0.ggpht.com/news/tbn/dNDwpcYNGpFYaM/0.jpg - DIRECT/74.125.153.103 image/jpeg
1282361594.561 288 171.11.238.157 TCP_MISS/200 1772 GET http://nt3.ggpht.com/news/tbn/N-iKYy_kXVXPuM/0.jpg - DIRECT/74.125.153.104 image/jpeg
1282411485.944 10 60.247.185.54 TCP_MEM_HIT/200 834 GET http://videos.asianbabemedia.com/jp18babydoll.asx - NONE/- video/x-ms-asf

Figure-4. A sample of proxy logs file.

4.2. Features Extraction. Web log file contains many attributes (fields). Only necessary fields are selected to be used in web user clustering while rest of attributes is dropped. The features extraction is a process of identifying, selecting and removing of unnecessary or irrelevant fields and/or rows form row log data. It includes two steps: data cleaning and extracting of the desired features.

Data cleaning involves the removal of irrelevant requests from the logs proxy files since some of the entries are not valid or not relevant. The data cleaning is carried out as follows:

- **Parsing:** This involves identifying the boundaries between successive records in log files as well as the distinct fields within each record.
- **Filtering:** This includes elimination of irrelevant entries and entries with unsuccessful HTTP status codes. The successful entries with 2xx and 3xx status codes are only considered in this paper.
- **Finalizing:** This involves removing unnecessary fields. Moreover, each unique user and URL should be converted to a unique integer identifier to ease the implementation.

In the step of extracting of the desired features, the desired features of web requests are extracted from the logs files. In this paper, we are interested in identifying the users and the common pages to know clusters of the users' interests. In order to know who visited the Web site, the log file must contain a person ID such as login to the server or to the user's own computer. However, most Web sites do not require users to log in, and most Web servers do not make a request to learn the user's login identity on her/his own computer. Thus, the information available according to the HTTP standard is not adequate to distinguish among users from the same host or proxy. More often it is an IP address assigned by an Internet Service Provider (ISP) or proxy server to a user's TCP/IP connection to the site, preventing unique identification. In addition to web users, the requesting web pages are extracted from logs files. So, entries for access of JPEG, GIF file, Java Scripts, other audio/video files need to be removed as they are executed or downloaded not on basis of user's request and hence might be redundantly recorded in log files.

4.3. Preparation of Training Dataset. As the features are extracted, these features must be prepared properly in order to obtain more accurate results. This step involves manipulating the dataset into a suitable form with training of the Fuzzy-c-means. So, base vector and training patterns preparation are required in this step.

The base vector consists of set of the preferred pages that can be utilized in knowing users' interests. The base vector B={ P1, P2…Pn } represents the access patterns of the users. If a web page is visited by three users or more it will be considered as one favourite page. Then, it will be included in the base vector B. For each user U, we form a user pattern vector PU, which is an instance of the base vector B. The pattern vector PU is formed by mapping the frequency of visits of each page in base vector B by user. Eventually, all the patterns of users are arranged to form the frequency matrix, in which the rows represent the users while the columns represent the preferred pages that form the base vector B. In addition to the integer patterns of users, binary patterns of users also can be used in order to discover users' interests in Web pages. In the binary patterns of users, contents of the pattern vector PU have value of either zero or one. Contents of the pattern vector PU

will be assigned to one if the Web page is requested by the users three or more times. Otherwise, It will be assigned to 0.

4.4. Web Users Clustering Using Fuzzy-C-Means. Once the dataset is prepared properly, the fuzzy-c-means clustering can be used depending on the finalized dataset for web user clustering. The fuzzy-c-means can cluster users by assigning the membership to each user pattern corresponding to each cluster centre on the basis of distance between the cluster centre and that user pattern. **Figure-5** shows the Web users clustering based on Fuzzy c-means.

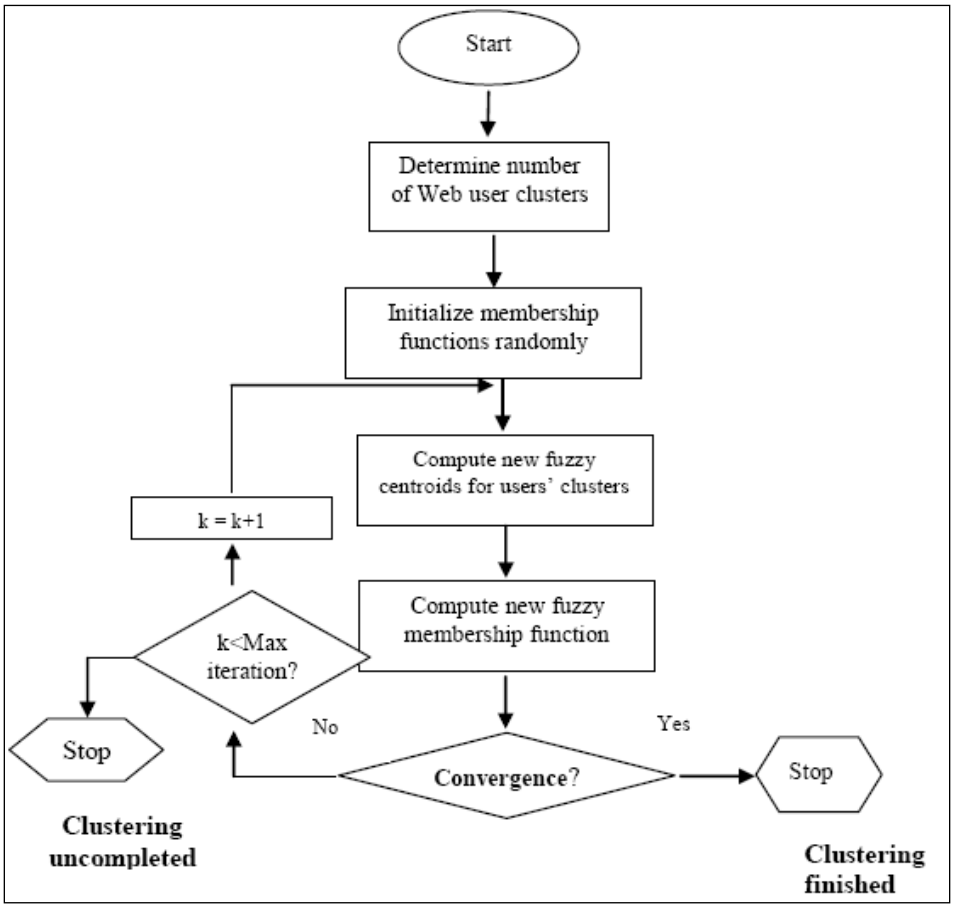

Figure-5: Web user clustering method based on Fuzzy c-means

The fuzzy-c-means begins with a decision on the number of web users clusters. Then, the memberships for web users of clusters are initialized randomly. At the k-th step, the fuzzy centroids are computed using equation 2. Subsequently, the fuzzy memberships are accordingly updated based on the new centroids of web users clusters as equation 3. If the change in membership functions|| U(k+1) - U(k)|| is zero or too small the convergence will be occurred. If convergence does not occurred and the number of iterations is less than maximum value of iterations then new fuzzy centroids and membership functions will be recalculated.

5. Results and Discussion. In this study, the raw datasets were obtained from the proxy logs files and traces of web objects requested in two proxy servers located around the United States of IRCache network [15]. The proxy logs files were collected from two proxy servers, namely: BO2 and NY. The BO2 and NY proxy logs files were used as datasets to evaluate the web user clustering techniques. In order to evaluate Fuzzy-c-means, we use equation (1) as performance objective function which is simplest and most widely used criterion measure for Fuzzy-c-means clustering.

In all our experiments, the performance objective function of Fuzzy-c-means decreases as the iterations number increases. If the objective function become stable and do not change convergence, the web users clustering will be successfully completed and Fuzzy-c-means will stops iteration. In all our experiments, the Web users clustering were successfully performed and very fast. Table 1 summarizes the parameters used in our experiments.

Table 1. Parameters of Fuzzy-c-means used in experiments

Parameter	Value
Clusters No.	2 - 6
Fuzzy parameter(m)	2
Max Iteration No.	100

5.1. Fuzzy-C-Mean Performance with Varying Clusters Number. Tables 2 and 3 present the performance of Fuzzy-c-means with varying clusters number of Bo2 and NY datasets for both real and binary forms. As can be seen, the Web users are successfully clustered using Fuzzy-c-means to similar groups very fast. As can be observed from Tables 2 and 3, objective functions of Fuzzy-c-means are reduced and become better when the clusters number increases for both of Bo2 and NY datasets. However, the performance Fuzzy-c-means with binary patterns of Bo2 and NY datasets is much better compared with the performance Fuzzy-c-means with real patterns of Bo2 and NY datasets.

Table 2. Performance of Fuzzy-c-means with varying clusters number for Bo2 dataset

	Binary Patterns		Real Patterns	
Clusters No.	Objective Function	Iteration No.	Objective Function	Iteration No.
2	16.96	12	4851.44	11
3	10.93	56	776.28	12
4	7.53	33	76.92	11
5	5.48	39	35.57	25
6	4.28	28	19.01	19

Table 3. Performance of Fuzzy-c-means with varying clusters number for NY dataset

	Binary Patterns		Real Patterns	
Clusters No.	Objective Function	Iteration No.	Objective Function	Iteration No.
2	31.79	17	8694.98	13
3	20.29	21	4170.03	26
4	14.68	79	2466.29	20
5	11.49	26	567.98	12
6	9.42	26	280.06	34

5.2. Comparison of Fuzzy-C-Means with K-means. From our experiments, It can be observed that K-means sometimes does not work well because It create empty clusters for web users. This is due to initializing centroids randomly that leads to some centroids starting close together. Some of these centroids end up becoming near-empty clusters, as a single cluster accumulates all of the data points. On the contrary, in Fuzzy-c-means, each web user has a degree of membership of belonging to each cluster. So Fuzzy-c-means always can cluster web users in a good way. In terms of computational time, we compare performance of K-means and Fuzzy-c-means in the web users clustering based on both real and binary patterns of Bo2 and NY datasets. Both K-means and Fuzzy-c-means performed the web users clustering very fast as shown in

Tables 4 and 5. However, Fuzzy-c-means sometimes was slightly slower than K-means since Fuzzy-c-means was actually doing more work. Every web user was assigned to each cluster, and so many operations were involved in each evaluation.

Table 4. Comparison of time(in seconds) between K-means and Fuzzy-c-means for Bo2 dataset

	Binary Patterns		**Real Patterns**	
Clusters No.	**K-means**	**Fuzzy-c-means**	**K-means**	**Fuzzy-c-means**
2	0.099875	0.007941	0.115619	0.009225
3	0.098960	0.309978	0.116297	0.008505
4	0.098359	0.122752	0.119538	0.008266
5	0.099784	0.021134	0.119279	0.009887
6	0.105263	0.012983	1.216373	0.010827

Table 5. Comparison of time(in seconds) between K-means and Fuzzy-C-Means for NY dataset

	Binary Patterns		**Real Patterns**	
Clusters No.	**K-means**	**Fuzzy-c-means**	**K-means**	**Fuzzy-c-means**
2	0.098715	0.009757	0.094540	0.007840
3	0.099750	0.009324	0.096407	0.010328
4	0.093280	0.018867	0.121189	0.009179
5	0.111953	0.012098	0.121764	0.008700
6	0.175808	0.875632	0.128975	0.014583

6. Conclusion. In this paper, Fuzzy-c-means clustering was used to cluster the Web users and discover users' interests based on Web pages. Unlike the conventional clustering methods, the proposed Web users clustering method based on Fuzzy c-means produced a degree of membership of Web user for every cluster. The experimental results depicted that the Web users were successfully clustered to similar groups very fast using Fuzzy-c-means. In addition, the Fuzzy-c-means performed well with varying clusters number on two real Bo2 and NY datasets. The proposed Web users clustering method based on Fuzzy-c-means can contribute in enhancing many applications such as Business Intelligence, E-Commerce, Customer Relationship Management, User Profiling and Personalization, Web site Construction, Fraud Detection, and Web Cache Replacement and Prediction.

REFERENCES

[1] Koutsonikola, V. A., Vakali, A. I. (2009). A fuzzy bi-clustering approach to correlate web users and pages. *International Journal of Knowledge and Web Intelligence*, 1(1-2), 3-23.

[2] Liu, B. (2007). *Web Data Mining: Exploring Hyperlinks, Contents*, and Usage Data. Springer Verlag

[3] Eirinaki, M., Vazirgiannis, M. (2003). Web mining for web personalization. *ACM Transactions on Internet Technology*, 3(1), 1–27.

[4] Frémal, S., Lecron, F. (2017). Weighting strategies for a recommender system using item clustering based on genres. *Expert Systems with Applications*, 77, 105-113.

[5] Neelima, G., & Rodda, S. (2016). Predicting user behavior through sessions using the web log mining.

2016 International Conference on Advances in Human Machine Interaction (HMI) (pp. 1-5). IEEE.

[6] Maged Mohammed (2016). *Web users clustering by using k-means and bottom-up algorithms*: Master Thesis. Institute of Science, Banaras Hindu University, India.

[7] Xu, J., & Liu, H. (2010). Web user clustering analysis based on K-Means algorithm. *In 2010 International Conference on Information*, Networking and Automation (ICINA).

[8] Feng, W., Kazi, T. H., Hu, G. (2012). Web Prefetching by ART1 Neural Network. *In Software and Network Engineering (pp. 29-40)*. Springer Berlin Heidelberg.

[9] Rangarajan, S. K., Phoha, V., Balagani, K., Selmic, R. R., Iyengar, S. S. (2004). *Web user clustering and its application to prefetching using ART neural networks*. IEEE Computer, 45-62.

[10] Chimphlee, S., Salim, N., Ngadiman, M. S., Chimphlee, W., Srinoy, S. (2006). Rough Sets Clustering and Markov model for Web Access Prediction. *In Proceedings of the Postgraduate Annual Research Seminar (pp. 470-475)*.

[11] Vakali, A., Pallis, G., Angelis, L. (2007). *Clustering Web Information Sources*.

[12] Bezdek, J. C.(1981). Pattern Recognition with Fuzzy Objective Function Algorithms. *New York: Plenum Press*.

[13] Zhang, L., Lu, W., Liu, X., Pedrycz, W., Zhong, C. (2016). Fuzzy c-means clustering of incomplete data based on probabilistic information granules of missing values. *Knowledge-Based Systems*, 99, 51-70.

[14] Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters*, 31(8), 651-666.

[15] NLANR (2010) National Lab of Applied Network Research (NLANR). Sanitized access logs: http://www.ircache.net/